MsDetector Training Pipeline

The pipeline consists of a collection of Perl scripts to perform the training and the evaluation of MsDetector on a chromosome of interest to the user. The master script is drive.pl. The user needs to update the values of a few variables in this script. Once this pipeline is installed properly, it can be used to generate the HMM and the GLM specific to a species of interest. MsDetector uses these models to extract microsatellites (MSs) in other chromosomes.

Contact

If you have questions, feel free to contact the first author at girgishz@mail.nih.gov or hani.z.girgis@gmail.com

Dependencies

The pipeline requires Perl, Matlab, and the Netlab toolbox. The Netlab toolbox can be downloaded from: http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/

Additional Manual

The manual distributed with the executable should be read first.

Variables

Next, we explain the variables requiring updates and their values.

$chromFile

A file includes the training chromosome in FASTA format. The header form is >chromosome:start-end. Use the fully-qualified name of the file.

Example:

```
>chrIV:0-1531933
ACACCACACCCACACCACACCCACACACACCACACCCACACACCACACCC
ACACCCACACACCCACACCCACACACCACACACCACACCACACCACACCC
ACACCCACACCACACCCACACCCACACACCACACCCACACCCACACACCA
CACACTACCCCTAACACTACCCTATTCTAACCCTGATTTTACCTGTCTCC
AAACCTACCCTCACATTACCCTACCTCCCCACTCGTTACCCTGCCCCACT
...
```

$repeatsFile

The fully-qualified name of a file containing the MSs, which can be obtained by scanning the training chromosome by RepeatMasker. Each line of this file includes a genomic location of the form >chromosome:start-end.

Example:

```
chrIV:4-155
chrIV:20525-20562
chrIV:33322-33353
chrIV:54130-54150
chrIV:56973-57028
```

Note

The chromosome name must be the same in the training chromosome file, $chromFile, and in the MSs file, $repeatsFile. For example, the chromosome name in the above examples, chrIV, is the same.

The coordinates system

Coordinates are zero-based. The start coordinate is inclusive, whereas the end coordinate is exclusive. Consequently, length = end - start. Use this convention to specify a header of the form >chromosome:start-end. The locations of RepeatMasker MSs and those located by MsDetector follow this convention as well.

$dstDir

This is the working directory where all intermediate files, subdirectories, and models are kept.

$freqFile

A file includes the nucleotide probabilities of the genome of interest. This file is needed to calculate the composition-correction scoring matrix. This file has one line which includes the probabilities of A, C, G, and T in this specific order and separated by a space(s). These probabilities can be calculated by the program NucleotideFreqMaker[32/64/Mac]. This program requires a directory including the chromosomes comprising the genome of the species of interest. The sequence of a chromosome must be in FASTA format. Chromosome files under this directory must have '.fa' extension. Each chromosome file must include one sequence only.

$chromShuffledFile and $useShuffledChrom

If the user wishes to use a special file to calculate the false positive rates, then the value of $chromShuffledFile is the user-specified file and the value of $useShuffledChrom is 1. Otherwise, the values of $chromShuffledFile and $useShuffledChrom are undef and 0, respectively; in this case, the false positive rates are calculated according to the shuffled version of the training chromosome. A zero-order Markov model is assumed while shuffling the training chromosome. Note that the length of the shuffled chromosome has to be the same as the length of the training chromosome in order to calculate the precision correctly.

$canPrepare

If the value of this variable is 1, the program prepares the required files including the three partitions for training, validation, and testing and the corresponding three shuffled sequences. Once this step is performed, the value of this variable can be assigned 0 to speed the subsequent executions.

$len

The value of this variable is the length of the motif, e.g. 6.

@setList

This list includes the names of the partitions used for training, validation, and testing, e.g. ('train', 'query', 'test'). The user may use the training set alone or in addition to any of the other two sets.

@factorList

The length of one of the two flanking sequences, i.e. a half window, is a multiple of the motif length. For example, if the motif length is 6, a factor of 4 results in a half window of size 24, e.g. (2, 4, 8, 12, 16).

@matrixList

This list includes the names of the scoring matrices that the user wishes to apply, e.g. ('Id', 'Trans', 'Comp', 'TransComp') representing the identity, the transition, the composition-correction, the transition & composition-correction scoring matrices.

$msDetector

The value of this variable is the fully-qualified name of the executable of MsDetector, not the optimized version.

$codeDir

The value of this variable is the fully-qualified path to the directory where the scripts comprising the pipeline are residing.

$matlab

The value of this variable is the fully-qualified name of the executable of Matlab.

$netlabDir

The value of this variable is the fully-qualified path to the directory including the '.m' files of the Netlab toolbox.

Output

In this example, the values of @matrixList, @factorList, and @setList were ('Id'), (4), ('train', 'query', 'test'), respectively. The following is the training, the validation, and the testing results. We refer to the validation set as the query set.

Results summary

8. Evaluating HMM ...
Id

x4 - train: Overlap length: 1741	Overlap: 90.54%	FP length = 2246	FPR = 4398.36755156234	Precision = 43.67%
x4 - query: Overlap length: 1304	Overlap: 87.93%	FP length = 2270	FPR = 4445.36702673487	Precision = 36.49%
x4 - test: Overlap length: 1003	Overlap: 90.61%	FP length = 2322	FPR = 4547.19031812707	Precision = 30.17%

10. Evaluating HMM+GLM ...
Id

x4 - train: Overlap length: 1659	Overlap: 86.27%	FP length = 48	FPR = 93.9989503450545	Precision = 97.19%
x4 - query: Overlap length: 1206	Overlap: 81.32%	FP length = 21	FPR = 41.1245407759613	Precision = 98.29%
x4 - test: Overlap length: 978	Overlap: 88.35%	FP length = 174	FPR = 340.745527714949	Precision = 84.90%

Overlap is the total length of the overlaps between RepeatMasker MSs and those located by MsDetector. In other words, it is the sensitivity to RepeatMasker detection. FP stands for false positive, and FPR stands for false positive rate. See the manuscript for the definitions of the sensitivity, the FPR, and the precision.

The Resulting HMM and GLM

The HMM and the GLM are written to hmm.txt and glm.txt under a subdirectory called Id4 which can be found under the working directory. The name of the subdirectory consists of the scoring matrix name and the window factor.

The GLM File

The glm.txt includes one line, e.g.

21.623 5.5018 24.097 0.1942 -1.7734 8.7013 2.8355

Numbers in this line are interpreted as the following:

The mean of the lengths of the +/- detections in the training set,
The standard deviation of the lengths of the +/- detections in the training set,
The mean of the average scores of the +/- detections in the training set,
The standard deviation of the average scores of the +/- detections in the training set,
The error or the bias,
The weight associated with the z-score of the length of a detection by the HMM, and
The weight associated with the z-score of the average score of a detection by the HMM.

The HMM File

The hmm.txt includes the prior, the transition, and the emission frequencies separated by empty lines. For example,

1016 5

508112 33
30 1304

0 0 226 72308 295746 125322 15444
0 0 0 12 94 380 968

The first line includes the prior frequencies of the series starting at the non-MS (S0) and the MS (S1) states. The counts of the transitions from S0 to S0 and S1 are listed in the third line. The counts of the transitions from S1 to S0 and S1 are listed in the fourth line. Lines 6 and 7 show the emission frequencies of the scores, 0-6, while the current state is S0 and S1, respectively. This file does not need further processing; MsDetector reads this file to generate the probabilities.